A Novel Feature Extraction Algorithm for Asymmetric Classification II

David Lindgren, Per Spångeus

Division of Automatic Control
Department of Electrical Engineering
Linköpings universitet, SE-581 83 Linköping, Sweden
WWW: http://www.control.isy.liu.se
E-mail: david@isy.liu.se

October 03, 2003

Report no.: LiTH-ISY-R-2550
Submitted to the IEEE Sensors Journal

Technical reports from the Control & Communication group in Linköping are available at http://www.control.isy.liu.se/publications.
A Novel Feature Extraction Algorithm for Asymmetric Classification
David Lindgren, Per Spångeus
3rd October 2003

Abstract
A linear feature extraction technique for asymmetric distributions is introduced, the asymmetric class projection (ACP). By asymmetric classification is understood discrimination among distributions with different covariance matrices. Two distributions with unequal covariance matrices do not in general have a symmetry plane, a fact that makes the analysis more difficult compared to the symmetric case. The ACP is similar to linear discriminant analysis (LDA) in the respect that both aim at extracting discriminating features (linear combinations or projections) from many variables. However, the drawback of the well known LDA is the assumption of symmetric classes with separated centroids. The ACP, in contrast, works on (two) possibly concentric distributions with unequal covariance matrices. The ACP is tested on data from an array of semiconductor gas sensors with the purpose of distinguish bad grain from good.

1 Introduction
The background to this work is the development of signal processing techniques for an array of semiconductor gas sensors used for detection of bad grain. The signals from the sensors are recorded under an interval of time when a gas mixer switches from clean dry air to gas picked up immediately above a grain sample. The switching induces pulses in the signals that are slightly different in magnitude and shape for different samples. A systematic difference enables a signal processing program to learn how to make more or less accurate statements about unknown samples. The recorded signals due to one sample are stacked in a measurement vector. The dimensionality of this measurement vector, or measurement for short, depends on the number of sensors as well as on the time duration of the recording and the sampling rate of the analogue-to-digital converter. Typically a measurement has \( n = 1000 \) entries or more. The signal processing uses a mathematical model that up to some degree of accuracy explains these measurements for different samples. This model is partly determined by measurements from a carefully designed set of calibration samples for which the correct statement (good/bad...
(grain) is known beforehand. In the end, the task of the signal processing is to make own statements about unknown samples, and we are interested in how the measurements best are modelled in order to make this statement reliable.

One major obstacle in this process is the very large dimensionality of the measurements, which make numerical stability, computational complexity and visualization issues that need consideration. We chose to address these difficulties by reducing every measurement to a small set with linear combinations. In this paper we propose a technique to determine these linear combinations for the good/bad classification problem, which is closely related to what is known as the problem of asymmetric classification, see below. We call the technique the asymmetric class projection (ACP).

Although this work was inspired by sensor arrays, the results are general and not limited to sensor systems. The discussion will thus be held on a rather general level. For more on the sensor arrays specifically, see for instance [12]. For an overview of signal processing methods for sensors, see [3, 8].

1.1 Probabilistic Framework

We assign to the vector $x_i$ in $\mathbb{R}^n$ the collected time discrete signals of the sensors due to measurement $i$. Since we do not know exactly what we measure, and since the measurement probably is perturbed with random noise, we assume that $x_i$ is actually a sample of the random variable $x$. From $x_i$ we now wish to extract a particular feature $y_i$ with high accuracy. $y_i$ is also drawn from a random variable $y$, since the feature that should be predicted is unknown at the time the measurement is conducted. The feature could for instance be a concentration of some chemical (quantification), the selection of a category or class from a set of classes (classification), it could be a gas detection or it could be the distinction between good and bad quality. The signal processing that maps the measurement $x$ to the feature $y$ we denote $f(\cdot)$, and by high accuracy is meant that the magnitude of the residual $r$ in the (inverse) regression

$$ y = f(x) + r $$

should be small in general. A common measure for residual magnitude is the expected quadratic $E[r^2]$. In our probabilistic approach we assume that $x$ has a $y$-conditional distribution.

By classification we define the regression problem (1) where $y$ is a bit vector $(0, 1)^q$ with one entry for each of $q$ categories classes. Every possible outcome of $x$ is associated by $f(\cdot)$ with one distinct class $j$, $1 \leq j \leq q$, and further with the bit vector

$$ y = [0_1 \cdots 0_{j-1} \ 1_j \ 0_{j+1} \cdots 0_q]^T $$

in such a way that the general probability of misclassification or error rate $E[r^T r] / 2$ is as small as possible. The lowest possible error rate given the distribution of $x$ is called the Bayes error, a limit of uncertainty no deterministic program can fall short of. In classification there are $q$ distinct $y$-conditional or class-conditional distributions, one for each class.
1.2 Regression in a Subspace

Since the dimensionality $n$ of $x$ is very high and the entries probably correlated, the regression in (1) is often difficult to solve directly by least squares or by maximum likelihood methods, see [13]. It is also very difficult to visualize high-dimensional data points. As mentioned above, one technique to overcome those difficulties is to consider an approximative problem limited to a small set of $k \ll n$ linear combinations of $x$. The linear combinations we define by the rows of a $k$-by-$n$ matrix denoted $S$. A new $k$-dimensional random variable $x_S$ is thus calculated by the matrix-vector multiplication $x_S = Sx$. The modified regression is now

$$y = \tilde{f}(Sx) + \tilde{r}, \quad (3)$$

and the objective is to find the linear combinations $S$ and function $\tilde{f}(\cdot)$ that make the magnitude of the residual $\tilde{r}$ small. This is known as subspace regression since the regressor is reduced to a subspace here defined by the range space of the rows of $S$.

There are many well-known and robust techniques to find a subspace in which the actual linear or nonlinear regression can take place. One of the simplest one is to use a limited number of uncorrelated principal components chosen after an error-impact or a variance criterion (PCR), see [6]. QR-decomposition is a similar way to orthogonalize the problem, see [2]. One popular technique in chemometrics is the partial least squares algorithm (PLS), which also can be formulated as a subspace regression method, see [10], and used for classification problems, see [11, 5].

Particularly for classification problems, the Fisher linear discriminant analysis (LDA), see [4] etc, is a technique to find linear combinations or discriminants that facilitates the ability to discriminate among $q$ distributions. As a measure of this ability is used the Fisher ratio,

$$\Delta = \sqrt{\frac{\sum_{j=1}^{q}(\mu_j - \mu)^T\Sigma^{-1}(\mu_j - \mu)}}, \quad (4)$$

where $\Sigma$ is the covariance matrix, assumed to be invertible and equal for all class-conditional distributions ($\Sigma_j = \Sigma$), $\mu_j$ the mean of class-conditional distribution $j$, and $\mu = E[x]$. The operator $E[\cdot]$ denotes expectation. In words, $\Delta$ is the amount of variation between distribution means with respect to the variation within the distributions, a measure maximized in the subspace given by LDA. Later on, LDA will be used as a reference technique.

1.3 The Asymmetric Classification Problem

The asymmetric classification problem concerns discrimination between (two) distributions with unequal covariance matrices. When the covariance matrices of the classes are equal, the optimal decision boundary to separate two normal distributions (minimal Bayes error) is given by the distribution’s symmetry plane, and it is not difficult to calculate the Bayes error, nor to find subspaces in which the Bayes error is minimal.

When the covariance matrices are unequal, it exists no symmetry plane, and the classes are said to be asymmetric. Still, the optimal classifier (which gives
Figure 1: An ACP of a data set with good measurements (rings) and bad (crosses).

error rate equal to the Bayes error) is well-known for normal distributions (the quadratic classifier), but it is much more difficult to calculate the Bayes error, especially in the multivariate case, see [7, ch. 3].

The typical example of an asymmetric problem that we will focus on, is when the two classes are of the type, good/bad, normal/abnormal, accept/reject. Say, for instance, that a sensor system should be used at a dairy and be “trained” to detect bad milk. The objective of the signal processing is then to transform a potentially high dimensional measurement into a binary output; good or bad. The distribution mean of the good class is not necessarily unequal to that of the bad class, but it is assumed that good can be defined as a restricted region in the feature space while bad is everything that does not lie in this region, see Figure 1. In this work we shall actually adopt the notion of a good class and a bad class, although the theory is general for asymmetric problems.

This work deals with the problem of how to find the best subspace $S$ for asymmetric classification. The subspace can be used to compress data, to enhance the prediction or just to visualize asymmetric classes in a 2-dimensional plot.

1.4 Reader’s Guide

In the next section, the ACP will be introduced and the objective explained as well as how to compute the ACP. After that, a result on Bayes error optimality that concerns the ACP will be presented. In the 4th section the ACP, Principal Component Analysis (PCA), and LDA will be compared when applied on a known artificial data set. In section 5, a numerical example with real data from a sensor array is subjected to ACP. Conclusions and acknowledgments are found last.
2 The ACP

As indicated above, the purpose of the ACP is to find a set of linear combinations or a linear subspace that contains as much relevant information as possible. This can alternatively be viewed as data compression, where we aim at retaining the ability to distinguish good from bad in the compressed data. The reduction from \( n \) to \( k \) dimensions (\( k < n \)) is defined by a set of \( k \) linear combinations, each of \( n \) variables, comprised in a \( k \)-by-\( n \) matrix \( S \). Compression of the measurement \( x_i \) in \( \mathbb{R}^n \) is calculated by the matrix multiplication \( x_i, S = Sx_i \). The measurement is thus projected onto the rows of \( S \), and we shall therefore denote this reduction by projection (although \( S \) is not a projection matrix in mathematical sense).

The mean of the projection \( x_S \) is calculated as \( \mu_S = ES \), and the covariance matrix as \( \Sigma_S = ESS^T \). In particular, if \( S \) is a row vector, both the mean vector and covariance matrix will be compressed to scalars with the mean and variance, respectively.

2.1 Objective

The fundamental assumption of the ACP is the existence of an ideal point in feature space. The degree of good will decay as we move away from that point. Thus, if we have two sets of data, one good and one bad, the measurements in the good set will be well clustered around the ideal point while the bad measurements will be more scattered and distant to the ideal point, see Figure 1. This is the basic property a classifier would exploit, and the property the ACP tries to retain in a projection.

Now, consider two \( n \)-dimensional random variables: \( g \) (the good) and \( b \) (the bad) with mean vectors

\[
\mu_g = \mathbb{E}[g], \quad \mu_b = \mathbb{E}[b],
\]

and covariance matrices

\[
\Sigma_g = \mathbb{E}[(g - \mu_g)(g - \mu_g)^T], \quad \Sigma_b = \mathbb{E}[(b - \mu_b)(b - \mu_b)^T].
\]

The distributions of \( g \) and \( b \) corresponds to the class-conditional distributions mentioned earlier. We assume that \( \Sigma_g \) is invertible and that \( x_i \) is a sample of either \( g \) or \( b \) with a priori equal probability. For a particular 1-dimensional projection \( x_s = s^T x, s \neq 0 \), the quotient between bad class variance and good class variance is given by

\[
\xi(s) = \frac{\text{Var}[s^T b]}{\text{Var}[s^T g]} = \frac{s^T \Sigma_b s}{s^T \Sigma_g s}.
\]

This quotient is a quality measure of the projection onto \( s \). Projections with larger \( \xi \) is preferred, since they facilitate the distinction between good and bad, as will be shown later.

In fact, the maximization of (the Rayleigh quotient) \( \xi(s) \) is a well-known problem equivalent to the generalized eigenvalue problem, [2]. The generalized eigenvalue problem is to calculate the matrix \( E \) with eigenvectors \( e_i \) and the diagonal matrix \( D \) with eigenvalues \( \lambda_i, 1 \leq i \leq n \), such that

\[
\Sigma_b E = \Sigma_g ED,
\]
\( e_i^T e_i = 1 \) and \( e_i^T \Sigma e_j = e_i^T \Sigma e_j = 0 \) whenever \( i \neq j \). We assume that the eigenvalues are ordered so that \( \lambda_i \geq \lambda_{i+1} \). The eigenvector \( e_1 \) then solves

\[
s^* = \arg \max_s \frac{\text{Var} [s^T b]}{\text{Var} [s^T g]}
\]

and the eigenvalue \( \lambda_1 \) identifies the optimum value. The vector \( e_1 \) gives the linear combination of \( b \) with largest variance compared to the variance of the same linear combination of \( g \).

Thus, by solving the generalized eigenvalue problem (8) a subspace is obtained where the good class is well clustered, the bad class well scattered. That this probably is a very good subspace in a Bayes error sense will be shown in section 3. The solution to the generalized eigenvalue problem is very well known and numerically stable and fast algorithms exist, see [9].

2.2 Modified Covariance

The covariance matrix \( \Sigma_b \) is defined as

\[
\Sigma_b = E \left[ (b - \mu_b)(b - \mu_b)^T \right].
\]

This is the standard definition of a covariance matrix, which means that the magnitude of the covariance is a measure of the variation or spread with respect to the mean. However, if the good and bad class are not concentric (\( \mu_g \neq \mu_b \)) it is more interesting for our purposes to measure the spread of the bad class with respect to the good class mean rather than to the bad class mean itself. This can be achieved by in (8) replacing \( \Sigma_b \) with

\[
\tilde{\Sigma}_b = E \left[ (b - \mu_g)(b - \mu_g)^T \right].
\]

With this definition of bad class covariance, \( s^T \tilde{\Sigma}_b s \) is a measure of how well the projection onto \( s \) spread the bad class with respect to the good class mean. Note that \( \tilde{\Sigma}_b \) is actually not a covariance matrix.

2.3 Generalization to More than One Dimension

The quotient of the modified variance of the bad and the variance of the good distribution is

\[
\xi = \frac{E \left[ (b - \mu_b)^2 \right]}{E \left[ (g - \mu_g)^2 \right]} = \frac{\sigma_b^2}{\sigma_g^2} = E \left[ (b - \mu_g)\sigma_g^{-2}(b - \mu_g) \right].
\]

(10)

As described earlier, this is used as a measure of discrimination in 1 dimension. The generalization to more than 1 dimensions we define as

\[
\xi = E \left[ (b - \mu_g)^T \Sigma_g^{-1} (b - \mu_g) \right] = E \left[ \text{trace} \left( \Sigma_g^{-\frac{1}{2}} (b - \mu_g) (b - \mu_g)^T \Sigma_g^{-\frac{1}{2}} \right) \right] = \text{trace} \left[ \Sigma_g^{-\frac{1}{2}} E \left[ (b - \mu_g)(b - \mu_g)^T \right] \Sigma_g^{-\frac{1}{2}} \right] = \text{trace} \left[ \Sigma_g^{-\frac{1}{2}} \tilde{\Sigma}_b \Sigma_g^{-\frac{1}{2}} \right] = \text{trace} \left[ \Sigma_g^{-1} \tilde{\Sigma}_b \right].
\]

(11)
Now, let $E$ and $D$ be the solution to the generalized eigenvalue problem,
\[
\Sigma_b E = \Sigma_g E D \quad \text{s.t.} \quad D, E^T \Sigma_g E \text{ and } E^T \Sigma_b E \text{ diagonal.} \tag{12}
\]
It is assumed that the eigenvalues on the diagonal of $D$ are ordered, $\lambda_i \geq \lambda_{i+1}$. The diagonality implies that the linear transformation $E^T g$ has uncorrelated components. If the distributions are concentric ($\mu_b = \mu_g \Rightarrow \Sigma_b = \Sigma_g$) this holds also for the components of $E^T b$. The diagonality also gives an easy way to calculate the trace,
\[
\xi = \text{trace} \left[ \Sigma_g^{-1} \Sigma_b \right] = \text{trace} [E D E^T] = \text{trace} [D] = \sum_{i=1}^{n} \lambda_i, \tag{13}
\]
since $e_i^T e_i = 1$. Thus, if we should chose among components that are uncorrelated with respect to $g$ and $b$, it is evident that we should take the $k$ ones with greatest eigenvalues. Thus, we define the $k$-dimensional ACP-projection by
\[
S_{ACP} = \left[ e_1 \ e_2 \ \cdots \ e_k \right]^T, \tag{14}
\]
and for this projection the discrimination measure is
\[
\xi(S_{ACP}) = \sum_{i=1}^{k} \lambda_i. \tag{15}
\]

### 2.4 Relation to LDA

As mentioned in Section 1, LDA finds the subspace with maximum
\[
\Delta^2 = \sum_{j=1}^{q} (\mu_j - \mu)^T \Sigma^{-1} (\mu_j - \mu). \tag{16}
\]
For two random variables (classes) $g$ and $b$ with equal a priori likelihood, and with equal covariance $\Sigma_g = \Sigma_b = \Sigma$, straight-forward calculation gives
\[
\Delta^2 = 0.5 \cdot (\mu_b - \mu_g)^T \Sigma^{-1} (\mu_b - \mu_g)
= 0.5 \cdot E [b - \mu_g]^T \Sigma^{-1} E [b - \mu_g] \tag{17}
\]
which is known as the (half, square) Mahalanobis distance between the classes. This should be compared to what is maximized by the ACP:
\[
\xi = E \left[(b - \mu_g)^T \Sigma_g^{-1} (b - \mu_g)\right]. \tag{18}
\]
It is seen that the major difference is that the expectation for the ACP is quadratic, $\xi = E \left[m^T m\right]$ for $m = (b - \mu_g) \Sigma_g^{-\frac{1}{2}}$, while for LDA it is $\Delta^2 = E \left[m^T m\right]$ for $m = (b - \mu_g) \Sigma^{-\frac{1}{2}}$, which of course is fundamentally different. Our interpretation of this is that the ACP maximizes a measure of variance while the LDA maximizes a measure of mean difference. LDA is not even defined when there is no mean difference.

Although both the LDA and ACP can be expressed as generalized eigenvalue problems, the solutions are different. For instance, (17) attain its maximum for the subspace dimensionality $k = 1$. Nothing is gained by taking $k = 2$ if we use (17) as measure. Since the rank of $\Sigma_g^{-1} \Sigma_b$ in (13) is very likely to exceed one, which corresponds to more than one non-zero $\lambda_i$, the ACP can however find more than one interesting dimension.
Thus, if $|x| > a$, $x$ most likely belongs to $b$.

3 Bayes Error

It will be shown that the projection onto $s$ is Bayes error optimal if $s$ is a solution to (9). The assumptions are that both classes are normally distributed with the same mean (concentric) and that $0 < \text{Var} \left[ s^T g \right] < \text{Var} \left[ s^T b \right] \forall s \neq 0$, that is, for every linear combination, the variance of the bad class is larger than the variance of the good class (which is greater than zero). To start with, two scalar random variables $g$ and $b$ are considered that are standardized so that $\text{Var}[g] = 1$ and $E[g] = E[b] = 0$. Thus, let $\xi$ be the variance of the bad class, $\xi > 1$.

Figure 2 depicts the probability density function (PDF) of $g$ and $b$ (standardized with respect to $g$). The normal PDF with zero mean and variance $\sigma$ is

$$p(x, \sigma) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sigma \sqrt{2\pi}}. \quad (19)$$

For decision boundaries $\pm a$, the error rate is given by the gray area in Figure 2,

$$\varepsilon(a, \xi) = \frac{1}{2} \left[ 2Q(a) \right] + \frac{1}{2} \left[ 1 - 2Q(a/\xi) \right] \quad (20)$$

where

$$Q(x) = \int_{-\infty}^{x} p(t, 1) \, dt. \quad (21)$$

The factors $\frac{1}{2}$ in (20) is the a priori probability of either class, which we have assumed to be equal for good and bad. By Figure 2 one realizes that the optimal decision boundaries ($\pm a$) for classification are given by the intersections...
between \( p(x, \xi) \) and \( p(x, 1) \) (they minimize \( \varepsilon(x, \xi) \): by moving the boundaries at \( \pm a \) in Figure 2, the sum of the gray areas can only become larger). Solving the equation \( p(x, \xi) = p(x, 1) \) gives the optimal boundary as

\[
a(\xi) = \sqrt{\frac{\ln \xi^2}{1 - 1/\xi^2}}.
\]  

(22)

A closed expression for the Bayes error thus comes out from (20) and (22) as

\[
\varepsilon(\xi) = \varepsilon(a(\xi), \xi) = Q\left(\sqrt{\frac{\ln \xi^2}{1 - 1/\xi^2}}\right) + \frac{1}{2} - Q\left(\sqrt{\frac{\ln \xi^2}{\xi^2 - 1}}\right).
\]  

(23)

Lemma 1. If measurements are a priori equally likely to be drawn from \( g \) in \( N(\mu, \sigma_g) \) as they are from \( b \) in \( N(\mu, \sigma_b) \) and \( \xi = \sigma_b/\sigma_g > 1, \sigma_g \neq 0 \), then the Bayes error decreases monotonically with the magnitude of \( \xi \).

Proof. With no change in the Bayes error, \( g \) and \( b \) can be affinely transformed (same transformation for \( g \) and \( b \)). No generality is thus lost by assuming \( \sigma_g = 1 \) and \( \mu = 0 \). It will now be shown that \( \varepsilon(\xi) \) as defined in (23) decreases monotonically when \( \xi > 1 \) increases. More specifically it will be shown that if \( \varepsilon \) is differentiated with respect to \( \xi \), the result is negative for every \( \xi > 1 \). Of course \( dQ(a)/da = -p(a, 1) \) and by the chain rule

\[
\frac{d\varepsilon}{d\xi} = -\frac{da}{d\xi} p(a, 1) + \frac{d(a/\xi)}{d\xi} p(a/\xi, 1).
\]  

(24)

Differentiating (22) gives

\[
\frac{da}{d\xi} = \frac{d}{d\xi} \sqrt{\frac{\ln \xi^2}{1 - 1/\xi^2}} = \frac{\xi}{D} \left(\xi^2 - 1 - \ln \xi^2\right)
\]  

(25)

and

\[
\frac{d(a/\xi)}{d\xi} = \frac{d}{d\xi} \sqrt{\frac{\ln \xi^2}{\xi^2 - 1}} = \frac{1}{D} \left(\xi^2 - 1 - \xi^2 \ln \xi^2\right).
\]  

(26)

\( D = \xi(\xi^2 - 1)^{\frac{3}{2}} \sqrt{\ln \xi^2} \) is a common denominator that apparently is positive for \( \xi > 1 \). Finally

\[
\frac{d\varepsilon}{d\xi} = -\frac{da}{d\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\ln \xi^2)}{\xi^2 - 1}} + \frac{d(a/\xi)}{d\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\ln \xi^2)}{\xi^2 - 1}}
\]

\[
= \frac{1}{\sqrt{2\pi}} \left( -\frac{da}{d\xi} \frac{\xi}{\xi^2 - 1} + \frac{d(a/\xi)}{d\xi} \frac{\xi}{\xi^2 - 1} \right)
\]

\[
= \frac{1}{D\sqrt{2\pi}} \xi^{-\frac{3}{\xi^2 - 1}} \ln \xi^2 (1 - \xi^2)
\]

< 0 if \( \xi > 1 \).
Theorem 1. If measurements are a priori equally likely to be drawn from \( g \) in \( N_n(\mu, \Sigma_g) \) as they are from \( b \) in \( N_n(\mu, \Sigma_b) \) and \( 0 < \text{Var}[s^T g] < \text{Var}[s^T b] \) for all non-zero \( s \) in \( \mathbb{R}^n \), then the vector

\[
\hat{s} = \arg \max_s \xi(s) = \arg \max_s \frac{\text{Var}[s^T b]}{\text{Var}[s^T g]}
\]

(27)
gives the Bayes error optimal projection to one dimension.

Proof. It will be shown that for two non-zero vectors \( s_1 \) and \( s_2 \) with \( \xi(s_1) > \xi(s_2) \), the Bayes error is smaller in the projection onto \( s_1 \) compared to the projection onto \( s_2 \).

It is well known that linear combinations of normally distributed variables are also normally distributed, thus \( s_i^T g \) and \( s_i^T b \) are normally distributed. Trivially \( s_i^T \mu = E[s_i^T g] = E[s_i^T b] \) (in every projection, \( g \) and \( b \) have the same mean). Furthermore, \( \text{Var}[s_i^T g] < \text{Var}[s_i^T b] \) why \( \xi(s_i) > 1 \forall s \neq 0 \). Since \( \xi(s_1) > \xi(s_2) \) the result follows directly from Lemma 1.

3.1 \( k \)-dimensional Projection

The explicit calculations of Bayes error optimality in multi-dimensions will not be developed in this work. It shall be pointed out, though, that the solution to the generalized eigenvalue problem gives components that are uncorrelated with respect to both \( g \) and \( b \), or equivalently, \( E^T \Sigma_g E \) and \( E^T \Sigma_b E \) are diagonal. The optimal dimensional extension to the principal eigenvector \( e_1 \) is thus \( [e_2 \cdots e_k] \) if the components (linear combinations) should be uncorrelated in the sense \( e_i^T \Sigma_g e_j = e_i \Sigma_b e_j = 0 \) whenever \( i \neq j \).

3.2 Non-Concentric Classes

Using the framework above, and the modified covariance explained in Section 2.2, it is necessary to show that for increasing values of \( \xi = \mu^2 + \sigma^2 \) the Bayes error can only decrease. Here the \( \mu \) and \( \sigma \) denote the mean and variance of the bad class assuming the variate has been standardized so that the good class has zero mean and unit variance. Numerical experiments indicate that the Bayes error decreases monotonically whenever \( \sigma^2 > \mu^2 \), but this remains to be shown analytically.

4 Artificial Example

The ACP of a computer generated data set shall be studied and compared to well-known and common techniques for feature extraction, namely PCA and LDA. The data set is not designed to mimic real life measurements, but rather to illustrate an instance where the ACP is superior.

Principal component analysis (PCA) is possibly the most common processing method in chemometrics. It is an unsupervised technique that concentrate as much variance as possible into few uncorrelated components. With unsupervised technique is understood a technique that does not utilize class labels (\( y \)). In other words, the knowledge of which measurement that are good and which
that are bad is not an input to PCA. LDA is a supervised technique that has been described earlier.

4.1 Artificial Data Set

The data set is originally 3-dimensional and 2-dimensional projections produced by PCA, LDA and ACP will be compared. The artificial data set describes two coaxial cylinders; the good class contained within the bad, 80 measurements in each class. Of course, the best discriminative projection in this case is a radial section of the cylinders. However, to fool the PCA much variance is given the data set in the axial direction. To fool the LDA a slight mean displacement is present, this also in the axial direction. The classes are thus not concentric.

4.2 Comparison

In Figure 3 the outcome of the different methods are compared. As intended, the PCA favors the direction with much variance and the projection is thus aligned with the axes of the cylinders. In this 2-dimensional PCA-subspace, a quadratic classifier has an (empirical) error rate of 11 misclassified measurements among a total of 160 measurements. Also the LDA favors the axes direction due to the mean difference. The error rate for the LDA is 15/160. The ACP is more or less a radial section that very well concentrates the good class in the center, and spreads the bad class around it. The error rate is 0.

5 Experimental Data

We study a data set obtained from measurements on 204 grain samples. A human test panel classifies each grain sample as good or bad. A sensor array with 23 response variables measures on the same samples.

Thus, for every measurement we have 23 variables and the knowledge whether it is attributable to a good or a bad grain sample. The entities \( \mu_g, \Sigma_g, \mu_b, \Sigma_b \) are unknown and have to be estimated from the data set itself. We now want to find out if the sensor configuration in the system can be trained to make a distinction between good and bad similar to the one produced by the human test panel. We also want to compare the feature extraction of PCA, LDA and ACP.

5.1 Validation

Since the means and covariances have to be estimated from the data set itself, random estimation/validation partitionings will be used to gain reliable results. The means, covariances and classification models are thus estimated on 75% of the measurements, and the displayed projection and classification accuracy are due to the remaining 25%. We denote the estimation sets \( T_{good} \) and \( T_{bad} \), respectively, and the validation sets \( V_{good} \) and \( V_{bad} \), respectively. The sets are described by matrices, where the columns are measurements. For instance,

\[
T_{good} = \begin{bmatrix} t_{1}^{good} & t_{2}^{good} & \cdots & t_{N_g}^{good} \end{bmatrix}
\]
Figure 3: The artificial data processed with PCA, LDA and ACP. Rings are *good* and crosses *bad*. The PCA selects the components with maximum variance, and thus loses the most relevant information needed to distinguish bad samples from good. The LDA does the same due to the mean difference direction, which in this data set is not optimal for discrimination. The ACP selects the cylinder radial section, which obviously is the best choice.
for the good class training measurements, where $t^\text{good}_i \in \mathbb{R}^{23}$. $N_g = 76$ is the number of measurements in the good class estimation set and $N_b = 77$ the number in the bad.

### 5.2 Estimation

The data set means $m^\text{good}$ and $m^\text{bad}$ are simply the arithmetic means of the respective training set. The data set covariance and modified covariance matrices $\Sigma^\text{good}$ and $\Sigma^\text{bad}$ are estimated as

$$C^\text{good} = \frac{1}{N_g - 1} (T^\text{good} - m^\text{good}1^T) (T^\text{good} - m^\text{good}1^T)^T$$

$$\tilde{C}^\text{bad} = \frac{1}{N_b - 1} (T^\text{bad} - m^\text{good}1^T) (T^\text{bad} - m^\text{good}1^T)^T,$$

where $1^T = [1 \ 1 \ 1 \ \cdots \ 1]$.

### 5.3 Projection

The optimal k-dimensional projection $S \in \mathbb{R}^{k \times 23}$ with respect to the quality measure trace $\left[C^{-1}_\text{good} \tilde{C}_\text{bad}\right]$ is calculated as (14), where $e_i$ are the principal eigenvectors of the generalized eigenvalue problem $\tilde{C}_\text{bad} E = C^\text{good} ED$. The validation data sets $V^\text{good}$ and $V^\text{bad}$ are projected as $V^S_\text{good} = SV^\text{good}$ and $V^S_\text{bad} = SV^\text{bad}$. They are plotted in Figure 6. See [9] for numerically stable algorithms for solving the generalized eigenvalue problem.

### 5.4 Scatter Plots

For a particular estimation/validation partitioning of the data set, scatter plots of the 2-dimensional projection of LDA and ACP are studied. As a reference, the plot of a PCA is depicted in Figure 4. It is seen, that in this subspace the accurate detection of bad samples is almost impossible. Comparing the plots of the LDA in Figure 5 to the ACP in Figure 6, one can see that the data have fundamentally different structure. LDA tries to find two separated clusters of the good and bad class, while the ACP centers around the good class, and tries to spread the bad measurements as much as possible. It is also seen that in both the ACP and the LDA subspace, distinction between good and bad can be done, although not very accurate. As mentioned earlier, for LDA on two classes it is only meaningful to study $k = 1$ dimensions; the vertical axis in Figure 5 displays only useless noise.

### 5.5 Empirical Error Rate

The error rate of a quadratic classifier in the 2 and 4-dimensional subspaces from PCA, LDA and ACP are given in Table 1. The figures are based on 100 random estimation/validation partitionings of the available data set, and for each partitioning, the subspace and classification model are calculated from estimation data, and the number of misclassified measurements of the 51 measurements in the validation data set is counted.
Figure 4: PCA of the grain validation data set.

Figure 5: LDA of the grain validation data set.
Table 1: Number of misclassified measurements out of 51 possible in different $k$-dimensional subspaces. The figures are the means $\pm$ standard deviations for 100 random estimation/validation partitionings of the grain dataset.

<table>
<thead>
<tr>
<th>$k$</th>
<th>PCA</th>
<th>LDA</th>
<th>ACP</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>23 $\pm$ 3.3</td>
<td>13 $\pm$ 3.2</td>
<td>14 $\pm$ 3.1</td>
</tr>
<tr>
<td>4</td>
<td>22 $\pm$ 3.3</td>
<td>14 $\pm$ 2.9</td>
<td>13 $\pm$ 2.6</td>
</tr>
</tbody>
</table>

As seen in the table, classification in the PCA subspace is not very much better than the flipping of a fair coin. Among the 2-dimensional subspaces, the LDA gives highest accuracy, and the ACP almost equally good. The opposite holds for the 4-dimensional subspace. The error rate in the unreduced 23 dimensions is 10 $\pm$ 2.9.

For this data set, there is no significant difference in the empirical error rate of LDA and ACP. This is probably because there is sufficient mean difference between the good and the bad class for the LDA to operate well. The Mahalanobis distance is estimated to about 2 standard deviations, and this distance gives a theoretical error rate of a *separating plane* of 15% or about 8/51 (*a priori* equally likely normal distributions with equal covariance matrices assumed). The grain data thus have a structure that is not the best for the ACP.

The sensor array can (in conjunction with the algorithms mentioned above) only up to a degree make the same distinction between good and bad as the human test panel. About 15% of the samples will be classified differently (if bad samples are as likely as good samples).
6 Conclusions

A method to find subspaces for asymmetric classification problems, the Asymmetric Class Projection (ACP), was introduced and compared to the well-known Linear Discriminant Analysis (LDA). The ACP has its main benefits when two distributions are nearly concentric and unequal in covariance. It was shown that for the concentric case (equal distribution means), the ACP is Bayes error optimal for at least 1-dimensional projections. The LDA cannot be used to analyze concentric distributions at all. An artificial data set showed an instance where one can expect the ACP to be beneficial to use.

Tested on a real data set, the LDA and ACP produced subspaces where the empirical error rates were about the same. The data set came from a gas sensor array that should detect bad grain samples. However, the structure of the ACP were in a sense more appealing: Measurements due to grain samples rated as good were well clustered around a fictitious ideal point, while measurements due to bad samples were more distant to the same point.

Acknowledgments

The ACP is a result of research at the Division of Automatic Control and the Swedish Sensor Center (S-SENCE). This work was partly sponsored by the Vetenskapsrådet (Swedish Research Council) and the Swedish Foundation for Strategic Research (the latter via the graduate school Forum Scientum). The contributions are gratefully acknowledged. We are also grateful to AppliedSensor AB for contributing with the grain dataset. The ACP algorithm (originally the GBP algorithm) as a calibration method for sensor systems is filed for a patent.

References


